CS592-LLM Project Final Report

Rajkumar Pujari
Purdue University
rpujari@purdue.edu

Abstract

In this project, we aim to investigate the usefulness of contextual information modeling in smaller Language Models such as Flan-T5 (Chung et al., 2022) for the '*Political Explanation Generation*' task. Political discourse tends to be notoriously vague, often intentionally so. The same text signals contrasting real-world actions depending on who says it and in which context. For example, in the sentence: "*We demand justice*", the word '*justice*' could mean completely different actions to people on opposing sides of the issue. Pujari et al. (2023) propose a '*Vague Text Disambiguation Generation*' task to test contextual understanding of NLP models. In this project, we design and evaluate a 4-step training process to improve performance on this task. We evaluate generated explanations using NLI-based proxy metrics on the full dataset and human evaluation on a subset of the dataset. We also evaluate the trained models on the classification variant of the task.

1 Introduction

Political discourse is often intentionally vague. Reasons for this include the rise of micro-blogging websites such as Twitter, the effectiveness of slogans and dog whistles, etc., Understanding such discourse requires an understanding of the context surrounding the text. That context includes the biases of the authors, details of the event in focus, the historical context of similar events related to the same issue in the past, etc. For example, the tweet in fig. 1 could be interpreted in completely different ways conditioned on the context surrounding it.

Pujari et al. (2023) proposes '*Vague Text Disambiguation*' task that tests the ability of NLP models to understand political text in its context. They show that large language models (LLMs) are outperformed by much smaller models when the context is modeled efficiently. Specifically, they adapt the Discourse Contextualization Framework (DCF) proposed by (Pujari and Goldwasser, 2021) to build a state-of-the-art model for the classification variant of the task. They also propose a generation variant of the task for which GPT-3 achieves 73.26% accuracy while a 20B parameter GPT-NeoX only reaches 20.04% even in a few-shot setting. This demonstrates that while smaller LLMs still lag on contextual understanding, even GPT-3 also lags significantly behind human performance on the task (>97%).

Figure 1: An example of varied *intended meanings* behind the same political message depending on the Author and Event in context

In this project, we train Flan-T5-base (250M parameters) (Chung et al., 2022) for the 'Vague Text Disambiguation' task. We propose a multi-task learning framework that trains the encoder portion of Flan-T5-* models using the learning tasks proposed for the DCF model in Pujari and Goldwasser (2021). We evaluate various training approaches to adopt Flan-T5 for this task.

First, we propose two multi-task learning paradigms to improve T5: (1) DCF model pre-training and (2) Politician quote pre-training. In stage 1, we train a multi-task architecture that

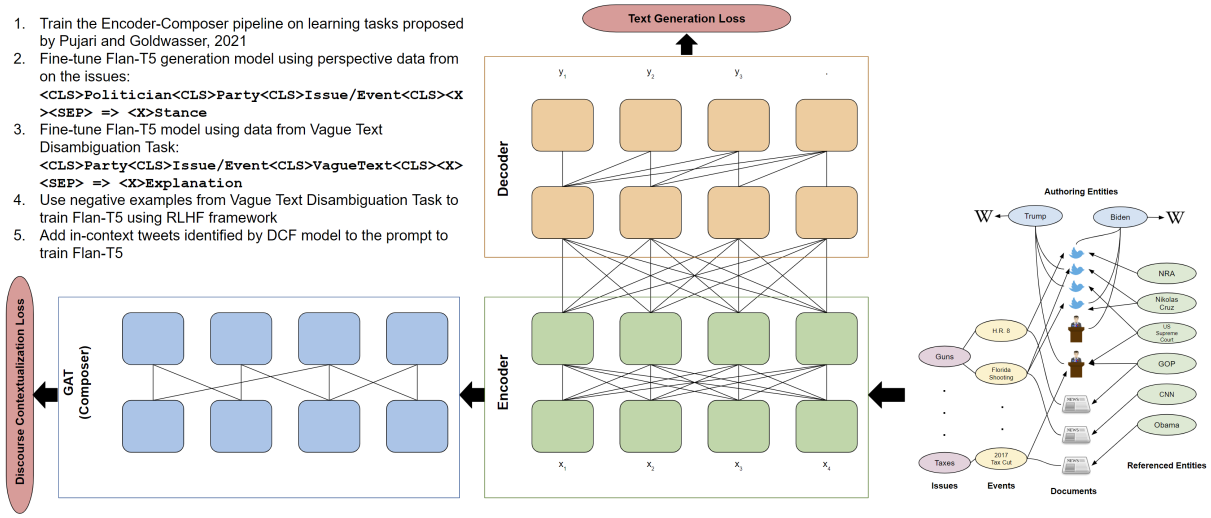

Figure 2: Flow Chart visualizing the proposed approach of our project. We skipped step 4 due to time constraints. Step 5 resulted in negative results.

shares the encoder part with T5 using the learning tasks proposed by Pujari and Goldwasser (2021). In stage 2, we train the T5 architecture using politicians' quotes on various issues such as Environment, Abortion, Guns, etc. Then, we fine-tune the T5 model on the classification and generation variants of the 'Vague Text Disambiguation Task'.

To evaluate the performance of the model on the generation variant of the task, we employ 2 different metrics: human evaluation and NLI-based proxy metrics. Inspired by Roit et al. (2023) and Honovich et al. (2021), we experiment with DeBERTa-large, a state-of-the-art model on MNLI dataset, to quantify the alignment of the generated explanations with the gold explanations. We observe that the existing NLI model is unable to accurately capture the correctness of the generated paraphrases. We experiment with fine-tuning the NLI model with in-domain data built from explanations from the 'Vague Text Disambiguation' dataset. For the classification variant of the task, we evaluate accuracy. We discuss the stages in more detail in section 3. We discuss the metrics in more detail in section 4.

2 Task Definition

The task of *Vague Text Disambiguation* evaluates the model's ability to identify a plausible explanation of an ambiguous quote given the event context and author affiliation. The rationale behind this task is that "*ambiguous language could be assigned grounded meaning when we know who is saying it and in which context*". For instance "*protecting our children from mass shootings*" could easily be disambiguated as either "*ban guns*" or "*arm teachers*" when we know the stance of the politician on the issue of '*gun rights*'. An example of the task is shown in table 1.

Vague Text Disambiguation Example
Vague Text: First, but not the last
Event: US withdraws from Paris climate agreement that enforces environmental targets after three years
Author Party: Republican
Disambiguation: The withdrawal from the Paris Climate Agreement is the first step of many to come for the Trump administration. It will not be the last, as more positive changes are sure to follow.
Incorrect Disambiguation: 1) Joe Biden's inauguration marks the first day of a new era of progress and prosperity, more lasting positive changes are coming. (Incorrect Event) 2) The Paris Climate Agreement withdrawal is the first of many backward steps this Trump administration is sure to take in destroying our environment. (Incorrect Stance) 3) This is the time for America to move forward and make progress without being held back by a global agreement that doesn't serve our interests. (Doesn't match the vague text)

Table 1: Example of *Vague Text Disambiguation Task*

3 Technical Approach

We train the Flan-T5 model in three stages: (1) DCF learning task training, (2) Politician quote generation training and (3) Task-specific fine-tuning. The original pipeline from the proposal, shown in figure 2 also included RLHF training and using in-context tweets from the DCF model. We couldn't perform RLHF training due to time constraints. We observed that adding in-context tweets from DCF model added noise to the input and affected the

performance of the model negatively. We discuss this further in section 4.

In stage 1, we train the encoder-composer architecture (horizontal stack in fig. 2) using learning tasks from Pujari and Goldwasser (2021). These tasks include Authorship Detection and Masked-Entity Detection tasks. We use the data released by Pujari and Goldwasser (2021) foe this pre-training. We share the encoder part of T5 as the text encoder for the task. This allows DCF learning tasks to backpropagate to the T5 architecture. As this changes the parameters of the encoder part of T5 without aligning with the decoder parameters, we don't evaluate this model directly on the 'Vague Text' task. Hence, we evaluate the performance of DCF pre-training by fine-tuning each variant of the task after DCF pre-training.

In stage 2, we train the T5-stack (vertical stack in fig. 2) using US politicians' quotes on various issues from `https://www.ontheissues.org/`. We use $27,775$ quotes from 24 political issues such as *Abortion, Civil Rights, Crime, Government Reform, Gun Control, Health Care, Immigration*, etc. We train the model in a seq-to-seq fashion. Our input prompt is formatted as `Issue [SEP] Party [SEP] Politician [SEP]`. We expect the model to generate the quote. We train the model based on loss from the generations. We evaluate this model directly on both variants of the task.

In stage 3, we perform fine-tuning on both variants of the tasks. We take the trained models from stage 1 and stage 2 and we fine-tune them for the classification and generation variant of the tasks. For the classification variant, we use the following prompt structure: `Instructions: Select the correct disambiguation of the vague tweet. Select choice 1, 2, 3, or 4. [SEP] Event [SEP] Party [SEP] Vague Tweet [SEP] Choices: (1) c1 (2) c2 (3) c3 (4) c4 [SEP] Answer:`. For the generation variant, we don't provide the choices in the prompt.

We further evaluate adding relevant tweets, identified by the DCF model, in the prompts. This resulted in noise being added to the prompt and consequent deterioration of results. A potentially better way of using the context graphs might be to let the representations computed by the DCF model be integrated into the decoder part of the generative model. This will allow both the T5 and DCF models to be trained on joint objectives as opposed to the multi-task learning setting proposed in this project. We leave this as a future work.

4 Evaluation

We evaluate the performance on the task of *'Vague Text Disambiguation Generation'* task using a combination of quantitative metrics and human evaluation. For human evaluation, we select a subset of 25 examples from the subset and we mark them as matching the metadata or not. We report the results in section 5.

For the NLI-based proxy metrics, we propose using the DeBERTa-large model trained on the MNLI dataset and classifying a generation as correct if it entails the gold generation (Metric 1). We observe that this metric doesn't capture the alignment between the generations and the gold explanations appropriately.

We devise an in-domain training task for the NLI model to adapt to the 'Vague Text' domain. We use the explanation choices from the dataset to create a training corpus for the NLI model. The dataset contains multiple plausible explanations for the same context (generated by different authors). These examples are essentially semantic equivalents as they are entailed by the same metadata. We use these as *entailment* data samples. We use the relation between correct and wrong choices of the same example as *contradiction* data samples. We use gold explanation choices from unrelated examples (different event and/or different vague text) as *neutral* data samples. We create a balanced dataset of 540 data samples to train the MNLI model. We report the generation performance with this model (Metric 2).

We also experimented with averaging the *entailment* scores of gold explanation and *contradiction* scores of negative explanation examples. But, as neural networks tend to be over-confident in their predictions (the probability of predicted class is always very high), the metrics correlated very highly with the above metrics. Hence, we don't report these metrics.

For the classification variant, we just report the accuracy on the test set. The 'Vague text Disambiguation" dataset contains 592 training examples and 145 test examples.

5 Results and Discussions

We report the results in table 2. T5 refers to the original LM. T5-Quotes is the model trained in stage 2 on politician quotes data. T5-FT is the original LM fine-tuned on the tasks. T5-DCF-FT is the model trained in stage 1 via multi-task learning

Model	Test Performance				
	Human	M1	M2	M3	Classification
T5	0	0.69	11.03	11.97	37.93
T5-Quotes	16	0.69	15.17	15.91	32.41
T5-FT	40	4.14	54.48	53.25	63.45
T5-DCF-FT	32	4.14	49.66	50.83	70.34
T5-Quotes-FT	48	1.38	46.21	46.66	64.83
T5-Quotes-DCF-FT	44	2.07	47.59	48.81	64.14

Table 2: Results of our Experiments on *Vague Text Disambiguation* Dataset Test Split

Vague Text: He voted his conscience. I know it's rare, but that's what happened.

Event: Second impeachment of Donald Trump ended with not guilty

Author Party: Democrat

T5: Trump voted his conscience, and he was not convicted.

Gold: It is a rare occurrence that one of the Republican senators chose to vote with his conscience and not follow along with the party line in the impeachment trial of President Donald Trump. Despite the rarity of this act, it is an admirable and honorable thing to do, and it is a sign of hope in a time of political turmoil

T5-Quotes-DCF-FT: While Trump was acquitted, there are a couple of republicans who made the right choice by voting their own conscience. It's nice to see these brave people.

Table 3: A Selected Example of Contextualized Generations

objective on DCF learning tasks. T5-DCF-Quotes-FT is the full model trained on both objectives and then fine-tuned on the 'Vague Text' dataset. We train on the DCF objective first and the Quotes objective second.

T5-Quotes-FT Generation Performance: We observe that T5-Quotes-FT performs best on the generation variant of the task according to the human evaluation metric. We also observe that the coherence of the generations is much higher in the T5-Quotes-FT model compared to other models. This shows the usefulness of unsupervised in-domain training on this task. This is not unfortunately not reflected in the automated metrics. This could either be attributed to the noisy metrics or lack of performance at scale. We couldn't investigate this further due to a lack of time and resources.

Complimentary Training Paradigms: It is interesting to note that the DCF pre-training is highly useful for the classification variant of the task whereas quotes pre-training is useful for the generation variant of the task. This shows that certain types of learning tasks and inductive bias is more suited a specific types of inference process. Hence, a unified model combining both paradigms, namely generative and discriminative paradigms, could be highly beneficial to the community.

Generation Error Analysis: While performing human evaluation, we note that the generation errors fall into 4 categories: irrelevant generations, incorrect party affiliation, incorrect event identification, and misalignment with the given vague text. Around 70% of the errors are irrelevant generations, as expected. The second largest portion of errors fall into incorrect party affiliation. This indicates that just providing the party name in the prompt might not be sufficient to fully align the generation 'persona'.

We also experiment with including tweets selected by the DCF model as context into the prompts directly. This results in a drop in generation quality as the LM is distracted by the tweet texts and generates unrelated text. This is an important negative result raising the need to design better integration pathways between the contextualization model and the generative LM.

Given more time, we would experiment with a more integrated contextualized LM. We would experiment with various joint training tasks to improve the performance of the contextualized LM.

6 Conclusion

In this project, we propose and investigate two training paradigms for the Flan-T5 model for two variant of political disambiguation generation task. We train the model in a multi-task setting using discourse contextualization learning tasks and political domain generative tasks. We observe that each of the training paradigms is useful for different inference paradigms. Interesting future work could be to extend the integration of discriminative and generative paradigms and design joint learning objectives that leverage the strengths of both paradigms in a complementary way.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rajkumar Pujari and Dan Goldwasser. 2021. Understanding politics via contextualized discourse processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1353–1367, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rajkumar Pujari, Chengfei Wu, and Dan Goldwasser. 2023. We demand justice!: Towards grounding political text in social context. In *Findings of EMNLP 2023*.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos Garea, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6252–6272, Toronto, Canada. Association for Computational Linguistics.